Identifying Vision Disorders Using Pupil Color Analysis

BY

Patrick G. Clark

Submitted to the graduate degree program in the
Department of Electrical Engineering and Computer Science
of the University of Kansas in partial fulfillment of the
requirements for the degree of Master's of Sciences.

_____
Chairperson: Dr. Arvin Agah

Committee members _____
Dr. Swapan Chakrabarti

_____
Dr. Jerzy Grzymala-Busse

Date defended: _____

The Thesis Committee for Patrick Clark certifies that this is the approved Version of the following thesis:

Identifying Vision Disorders Using Pupil Color Analysis

Committee:

_____
Chairperson: Dr. Arvin Agah

_____
Dr. Swapan Chakrabarti

_____
Dr. Jerzy Grzymala-Busse

Date approved: _____

Acknowledgements

I would like to thank my advisor, Dr. Arvin Agah. He has expended remarkable effort to make this a very interesting and thoroughly enjoyable project. Through solid guidance, insight, and just the right amount of assistance he has helped me to build a solid foundation in applied artificial intelligence.

Abstract

Amblyopia is a neurological vision disorder that studies show affects two to five percent of the population. Current methods of treatment produce the best visual outcome if the condition is identified early in the patient's life. Several early screening procedures are aimed at finding the condition while the patient is a child, including an automated vision screening system developed by Cibis, Wang, and Van Eenwyk. The system uses artificial intelligence software algorithms to achieve a 77% accuracy in identifying patients who are at risk for developing the amblyopic condition and should be referred to a specialist. This thesis aims to improve upon the existing automated vision screening system and increase the sensitivity, specificity, and accuracy measurements. It explores the application of decision tree learning algorithms and artificial neural networks on a previously unused set of features. The features are extracted from images of patient eyes and focus on the color information contained. The efficacy of pixel color data is also investigated with respect to the measurement of the rate of change of the color in the iris and pupil. Processing the data and testing the machine learning applications using a 10-fold stratified cross validation procedure reveals that the best results show an overall accuracy of 68% in identifying patients who are at risk of developing the amblyopic condition. These results do not outperform the previous research; however, the process has allowed an in-depth investigation into the potential of the iris and pupil color slope features.

Keywords: artificial neural network, decision tree, random forest, amblyopia

Table of Contents

Chapter 1: Introduction....1
1.1 Motivation....1
1.2 Research Approach....2
1.3 Problem Statement....4
Chapter 2: Background....5
2.1 Key Terms....5
2.1.1 Amblyopia....5
2.1.2 Anisometropia....6
2.1.3 Bruchner Reflex....6
2.1.4 Diopter....7
2.1.5 Foveation....7
2.1.6 Hirschberg Test....7
2.1.7 Hyperopia....9
2.1.8 Myopia....10
2.1.9 Photorefractive Screening....11
2.1.10 Strabismus....11
2.2 Analysis Methods....12
2.2.1 Holdout Testing....12
2.2.2 Confusion Matrix....13
2.2.3 Sensitivity, Specificity, and Accuracy....14
2.2.4 K-Fold Cross-Validation Testing....15
2.3 Amblyopic Risk Factor....15
2.4 Artificial Intelligence Techniques....16
2.4.1 Decision Trees....16
2.4.2 Random Forests....18
2.4.3 Artificial Neural Networks....20
Chapter 3: Related Work....23
3.1 Traditional Vision Screening....23
3.2 Photorefractive Screening....25
3.3 Automated Photorefractive Screening....27
3.4 Video Vision Development Assessment....28
3.5 Bruchner's Reflex....30
3.6 Automated Video Vision Development Assessment....30
Chapter 4: Methodology....33
4.1 Experiment Setup....33
4.1.1 Feature Extraction Overview....35
4.1.2 Processing Patient Videos....35
4.1.3 Extract Key Frame....36
4.1.4 Crop and Extract Eye Images....37
4.1.5 Extract Pixel Data....38
4.1.6 Produce Feature Groups....40
4.2 Classifier Training....43
4.2.1 J4.8 Decision Tree....44

4.2.2 Random Forest....45
4.2.3 ANN....46
Chapter 5: Experimental....48
5.1 Phase One....48
5.1.1 Raw Pixel....48
5.1.2 PCA Pixel....49
5.1.3 Four Color Slope....49
5.1.4 One Color Slope....50
5.1.5 Summary Results....50
5.2 Phase Two....51
5.2.1 Raw Pixel....51
5.2.2 One Color Slope....52
5.2.3 Summary Results....52
5.3 Phase Three....53
5.3.1 Raw Pixel....53
5.3.2 One Color Slope....54
5.3.3 Summary Results....54
5.4 Analysis....55
Chapter 6: Conclusion....57
6.1 Contributions....57
6.2 Limitations....57
6.3 Future Work....58
6.3.1 Key Frame Selection....58
6.3.2 Investigation of Foveating Frames....59
6.3.3 Include Color Slope with Previous Features....59
6.3.4 Additional Feature Investigation....60
6.3.5 Use Color Slope To Classify Diopter Error....61
6.3.6 Conclusion....62
References....63

List of Figures

Figure 2.1: Illustration of the Hirschberg reflex, point, and ratio.....9
Figure 2.2: A hyperopic (farsighted) eye, corrected with a convex lens.....10
Figure 2.3: A myopic (nearsighted) eye, corrected with a concave lens.....10
Figure 2.4: A confusion matrix.....14
Figure 2.5: Multi-layer perceptron.....20
Figure 2.6: Sigmoid Activation Function.....21
Figure 3.1: Snellen E Chart used in traditional vision screening.....24
Figure 3.2: VVDS System.....28
Figure 3.3: Key frame output from the AVVDA system.....31
Figure 4.1: Diagram of the feature extraction process.....35
Figure 4.2: An example of a key frame cropped around the pupils for each eye.....37
Figure 4.3: Pixel color extraction.....39
Figure 4.4: Phase two and three pixel color extraction.....40
Figure 4.5: An example of the color segments along the 45 degree axis.....43
Figure 4.6: Decision tree generated from the phase two one slope feature set.....45

List of Tables

Table 3.1 Accuracy results of the AVVDA system.....32
Table 5.1: Pixel value result detail.....48
Table 5.2: PCA Pixel value result detail.....49
Table 5.3: Four color slope value result detail.....50
Table 5.4: One color slope value result detail.....50
Table 5.5: Pixel value result summary.....50
Table 5.6: PCA Pixel value result summary.....51
Table 5.7: Four color slope value result summary.....51
Table 5.8: One color slope value result summary.....51
Table 5.9: Pixel value result detail.....52
Table 5.10: One color slope value result detail.....52
Table 5.11: Pixel value result summary.....52
Table 5.12: One color slope value result summary.....53
Table 5.13: Pixel value result detail.....53
Table 5.14: One color slope value result detail.....54
Table 5.15: Pixel value result summary.....54
Table 5.16: One color slope value result summary.....54
Table 6.17 Accuracy results of the AVVDA system and current research.....56

Chapter 1: Introduction

1.1 Motivation

Eye trouble of organic origin must be diagnosed and treated before the installation of an irreversible amblyopia or what is commonly referred to as “lazy eye”. This condition is a developmental disorder of the visual system caused by ocular abnormalities early in life. While surgery or optical correction of refractive errors can often address the initial cause of amblyopia, once amblyopia has developed, such interventions cannot restore visual function since amblyopia itself is a cortical deficit, a neurological disorder and not a physical one [Anderson 1999]; therefore, corrective action after amblyopia has developed becomes problematic as the brain will not be able to regenerate the neural pathways. Thus, early detection is essential for the patient to have a healthy visual outcome. Fortunately, amblyopia can be successfully treated if identified when the patient’s brain is still in the developmental stages, generally when the patient is fewer than six years old [Cibis 2005].

Amblyopia has two primary causes: strambismus and anisometropia. Strabismus is a misalignment between the two eyes. Anisometropia is when the refractive error between the two eyes is different [Steinman 2000]. The reason that these two conditions can lead to amblyopia is because they cause the brain to begin ignoring the signals from the weaker or blurrier eye. If left untreated, the neural paths degenerate and the weaker eye effectively becomes useless. If treated during the early development of the brain, the non-controversial methods of glasses and patching therapy achieve 80 to 90% effectiveness [Cibis 2005].

The most pervasive challenge to early diagnosis is enough specialists to screen patients at an early age to identify the condition. Since current methods of identifying the problem require well-trained operators or even medical doctors, the number of personnel available to evaluate even a small fraction of the worldwide population is not sufficient. According to some research, amblyopia affects from two to five percent of the population [Weber 2005][Robaei 2006]. Unfortunately a large percentage of the population lacks access to proper vision care to facilitate treatment of the problem in the optimal years (prior to six years of age) [Weber 2005]. In addition, the study by Weber and Woods shows that populations that undergo early intervention have a lower prevalence of amblyopia than those that do not [Weber 2005]. This implies the condition does not improve on its own accord and further supports the need for early accurate detection and treatment [Weber 2005].

1.2 Research Approach

The optimal solution for vision diagnosis would be a self-contained, low-cost, completely automated system that could accurately identify disorders with minimum operator training and patient cooperation. The solution would need to enable wide spread use with small children, including infants [Van Eenwyk 2008]. One idea for this sort of solution is based on the work of Gerhard W. Cibis, M. D. [Cibis 2005] [Wang 2002]. He pioneered the science of analyzing images for identifying features that may indicate the development of amblyopia. He then further advanced the project by teaming up with university researchers to develop an automated screening system using image processing and artificial intelligence techniques [Van Eenwyk 2008].

This project looks to build from the previous work on this system and analyze additional salient features extracted from the images. The overall goal is to use an automated screening system that can be used in locations without reliable access to vision care. It should be able to accurately identify children who are candidates for the amblyopic condition and who should be referred to a specialist. In addition, the screening process should be able to be handled by minimally trained operators and require minimal patient cooperation [Van Eenwyk 2008].

The approach to investigating this problem involves four steps. The first step is to capture the patient information that will be analyzed for any vision problems. The process involves an operator positioning the patient 52 inches from a video camera with a fiber optic light source mounted just below the optical lens. The second step is to record the patient looking at the light source for approximately two minutes. The purposes of the light source is to both capture the patient's attention and to reflect the light off the retina at the back of the patient's eyes so that the refraction through the patient's lens is captured. The reason that the patients must have their attention captured is because the age of the patient is typically less than six years old, people resistant to sitting still and focusing for two minutes. The third step is to process the video with a specialized software program that will analyze the patient video. The analysis of the video processes it into a set of frames that meet certain criteria for further image processing and feature extraction techniques. The fourth and final step is to send the features that were extracted through multiple pattern recognition algorithms in order to make a refer/no-refer decision that the operator can use. The overall goal of a system such as this is

to provide a very accurate referral mechanism so that only the patients who are at risk of developing serious vision disorders will be sent to a specialist.

1.3 Problem Statement

The specific problem to be studied centers on the methods through which amblyopia is detected. This project will explore the improvement of an existing automated system to identify vision disorders so that potential problems can be addressed as early as possible [Cibis 2005] [Wang 2002]. The existing automated system currently reviews 54 features that are extracted as a part of the image processing step. The features are further divided into 27 for each eye. Two features are considered of primary importance, the pupil radius and the degree of fixation based on the Hirschberg point. The remaining 25 features for each eye revolve around the color saturation in pupil region, thus focusing the attention of the refraction of the reflection from the retina [Van Eenwyk 2008]. The previous work will be more comprehensively discussed in chapter three.

The improvement to the current system will revolve around additional features that are to be extracted from the images. These features attempt to capture the rate of change of the pupil color along several axes. Dr. Cibis argues that some identifying markers in the color can accurately identify the amblyopic condition. The goal is that the additional features would yield more accurate results for a refer/no-refer recommendation when they are used in artificial intelligence classification algorithms.

Chapter 2: Background

2.1 Key Terms

This section provides a brief definition of key terms that are used in this paper. Understanding the medical concepts will allow the reader to better comprehend the research being presented.

2.1.1 Amblyopia

Amblyopia is the primary vision disorder the research presented attempts to accurately identify. As discussed previously it is a developmental disorder with roots in the physical structures of the eye. It has two primary physical causes, anisometropia and strabismus. Both of these physical abnormalities in the eye have the potential to cause the development of a patient's visual function to be impaired and cause the image from the amblyopic eye to be disregarded by the visual cortex. If the condition is allowed to persist, the neural pathways become permanently formed and the use of the amblyopic eye is diminished. The degree to which it is diminished varies based on how early the condition presented in the patient's life and if any remediation treatment was used [Steinman 2000]. When a patient displays the amblyopia markers, he or she is defined as an amblyope. In addition, when referring to the condition it is common to refer to the patient as having the amblyopic condition [Steinman 2000].

2.1.2 Anisometropia

Anisometropia is the medical term that is used to describe a condition where the refractive error in one eye is significantly different than the other. The definition is further qualified to be the cases where there is a difference in the sphere of the eye that is on diopter or more and no strabismus is present [Van Eenwyk 2008]. The difference in the refractive errors is difficult to overcome for a developing visual cortex primarily due to the very different images being presented. A particular study by Steinman, Steinman, and Garzia shows that anisometropia is the predisposing condition that leads to amblyopia 50% of the time and further studies find that an undiagnosed anisometropia will lead to strabismus [Steinman 2000].

2.1.3 Bruchner Reflex

The red reflex that occurs when a light source shines through the lens of the eye and reflects off the retina and then back to the observer is referred to as the Bruchner reflex [Cibis 1994]. It is named for its modern proponent who used an ophthalmoscope to view and measure the red reflex and compare the difference between the two eyes [Cibis 1994]. When the reflex is abnormal, the patient is considered to have positive Bruchner reflex with an abnormal Bruchner test. With a positive Bruchner test, the deviated eye is the brighter of the two [Cibis 1994]. Dr. Cibis explains this phenomenon as a result of the reflected light “leaking” into the observer's pupil rather than being focused directly back at the observer's light source [Cibis 1994]. Bruchner's reflex is a key feature that is used to as a marker to identify a physical abnormality of the eyes that could lead to the amblyopic condition.

2.1.4 Diopter

A diopter is a unit of measure used by the scientific community for describing the optical power of a lens. When used in the context of the human eye, scientists use the measure for describing the degree of focusing error, or refractive error, in the eye [Van Eenwyk 2008]. For a perfect eye, the refractive error is 0.00 diopters (D), while a hyperopic eye has a positive refractive error (e.g. +3.00 D) and a myopic eye will have negative refractive error (e.g. -1.50 D). Typically the refractive error is measured to the nearest quarter diopter [Van Eenwyk 2008].

2.1.5 Foveation

Foveation is a term coined by Dr. Cibis in his research on photo refractive screening methods. In the paper he published in 1994, foveation was used to identify the point of true fixation from a slightly off-axis fixation. Patients will sometimes experience an intermittent deviation of six degrees or less from true fixation [Cibis 1994]. An examination of recorded patient video reveals foveation when the focus goes from fixation to slightly off-axis fixation. The change will typically happen in one frame and is approximately 1/30 of a second [Van Eenwyk 2008]. Dr. Cibis argues that this sort of behavior may hold a large amount of information regarding the true vision of a patient.

2.1.6 Hirschberg Test

The Hirschberg test, or also called the Hirschberg cornea reflex test, is a screening test used in the fields of ophthalmology to identify strabismus. The procedure was pioneered by Julius Hirschberg in 1886 when he used a candle to observe the reflection of light through the

patient's corneas [Wheeler 1942]. In a patient with normal eye function, the reflection of the light from the cornea results in the center of the eye. However, those patients with abnormal eye function will reflect the light off center and the degree in which the reflection is off center can be measured to determine the degree of misalignment [Wheeler 1942]. The Hirschberg point is the point on the cornea that is the reflected light and a ratio is calculated based on that point and the edge of the pupil. The difference in the ratio is what determines the degree of misalignment between the two eyes [Cibis 1994]. Figure 2.1 illustrates the measurement of the feature. Today, the procedure is done with more sophisticated and sensitive tools, but the general idea presented by Hirschberg remains the same.

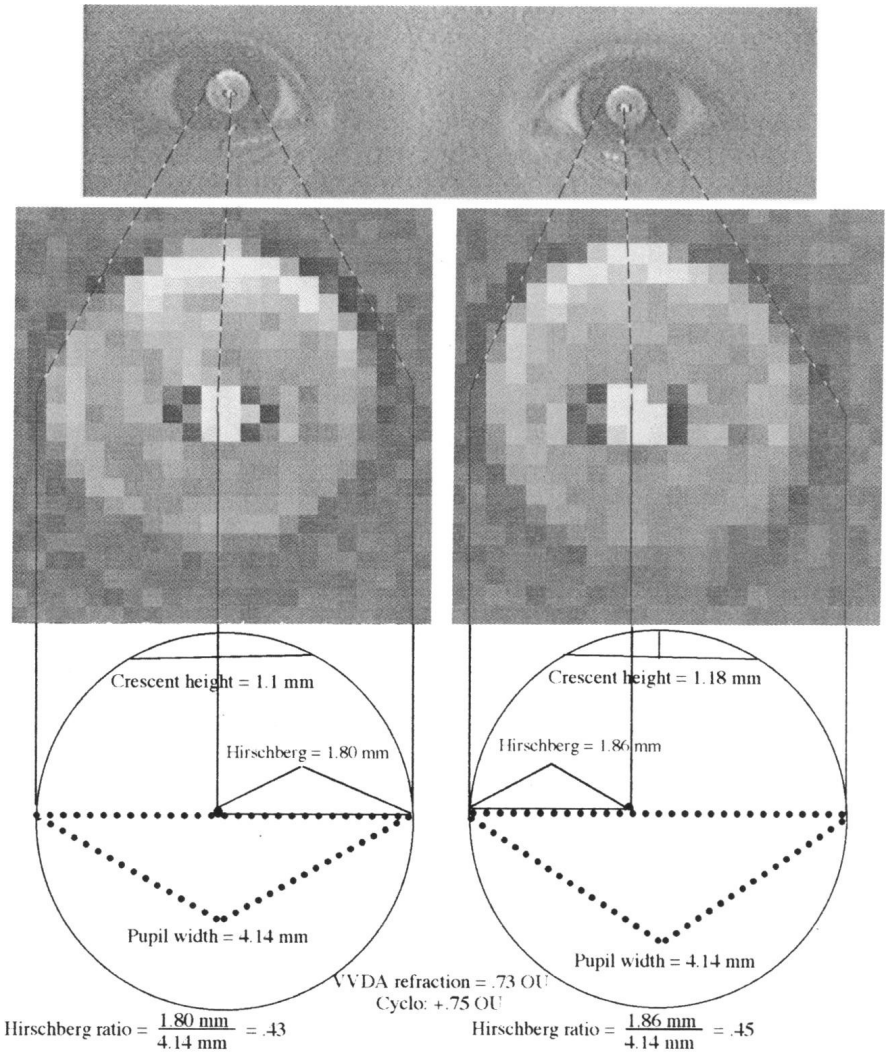


Figure 2.1: Illustration of the Hirschberg reflex, point, and ratio [Cibis 1994]

2.1.7 Hyperopia

Hyperopia is the medical term for a person who is farsighted. Medically speaking, a farsighted eye will focus the light through the lens behind the retina and will cause the image to appear blurry [Van Eenwyk 2008]. The hyperopic condition can be dealt with using a positive diopter lens, or a convex lens.

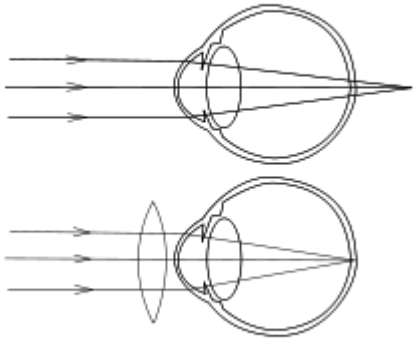


Figure 2.2: A hyperopic (farsighted) eye, corrected with a convex lens
Courtesy of Wikipedia Foundation, Inc.

2.1.8 Myopia

Myopia is the medical term for a person who is nearsighted. Medically speaking, a nearsighted eye will focus the light through the lens in front of the retina and will cause the image to appear blurry [Van Eenwyk 2008]. The myopic condition can be dealt with using a negative diopter lens, or a concave lens.

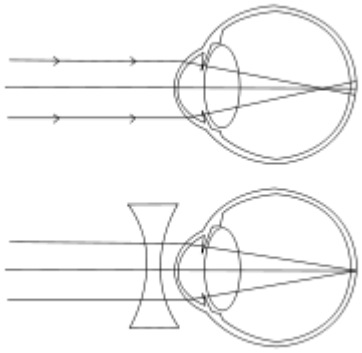


Figure 2.3: A myopic (nearsighted) eye, corrected with a concave lens
Courtesy of Wikipedia Foundation, Inc.

2.1.9 Photorefractive Screening

Photorefractive screening refers to the procedure used by ophthalmologists involving a light source and the non-linear optical effect seen as a result of shining the light source in the eye. Since the eye is essentially a transparent sphere, most of which is dedicated to focusing light on the back of the retina, how light travels through the lens and eye is essential to determining ocular health [Van Eenwyk 2008]. The goal is to capture the reflection of the light source off the eye in order to better understand the function of the patient's eye. The two types of photo refraction are on-axis and off-axis. Dr. Cibis chose to use the off-axis photo refraction techniques in order to highlight the crescent effect [Cibis 1994]. The degree and size of the crescent is important in identifying refractive errors between the two eyes (anisometropia) and any misalignment (strabismus). Using a recording device to capture the procedure allows the material to be reviewed in more detail by a specialist other than the person administering the test or, as in the case of this research paper, by a software system.

2.1.10 Strabismus

Strabismus is the second physical abnormality of the eyes that can lead to amblyopia. Commonly it can be identified as a misalignment of the focal point between the two eyes, but the condition is not always identifiable with the naked eye [Van Eenwyk 2008]. It may require a thorough screening before the condition presents. Strabismus typically involves a lack of coordination between the two eyes and the extra-ocular muscles where the patient is unable to bring both eyes into focus on the same point in space, thus preventing proper bi-ocular vision. A simple screening test is the Hirschberg test, where a light is reflected off the

patient's eyes. If the reflection is at the same place on both eyes, then the eyes are properly aligned [Steinman 2000].

2.2 Analysis Methods

In this section the techniques used for analysis of the artificial intelligence classifiers will be described.

2.2.1 Holdout Testing

When measuring the efficacy of classification algorithms for a particular dataset, the testing methodology used is very important. Choosing the correct method in which to train and test the classifier will affect its classification ability with future datasets that were not a part of the original data. In addition, a researcher should not train the system on data that will be used for testing; practicing this technique is analogous to giving the classifier the answers to the test. Holdout testing, sometimes called test sample validation, is a technique where the dataset is divided into two mutually exclusive subsets [Kohavi 1995]. One subset is called the training set and the other is the test set, or holdout set [Kohavi 1995]. Designating less than 1/3 the dataset for the holdout is common, and typically, the exact number to use is determined at the discretion of the researcher. Part of this discretion is applied when the assumption is made that a classifier's accuracy increases with the more instances it sees during training. For this reason, the holdout method is considered a pessimistic estimator because only a portion of the information is given to the classifier [Kohavi 1995]. The holdout data set is randomly selected from the entire dataset. Therefore, if the dataset that results from this selection is biased toward one particular pattern, it may negatively affect the

classifier's ability to learn all the patterns necessary for the most accurate classification [Kohavi 1995]. While using the holdout method has some drawbacks, it is still a valid technique for testing classifier performance. This sort of testing would be preferred in three instances. The first is when the testing computation time is long and precludes the use of a more thorough testing method. The second is for datasets that are large enough to provide good coverage of the classes that the classifier needs to learn. The final case is when the researcher is in the initial phases of research and desires a general idea of how a classifier may perform, a sort of initial indicator as to whether the research is valid to pursue [Kohavi 1995].

2.2.2 Confusion Matrix

The confusion matrix is a means by which to represent classification results in a tabular format and is the primary means this research paper uses to display classification results. Figure 2.4 shows an example confusion matrix for the data being used in this research experiment. The top row shows the summarized results of the “Refer” input class. In this case, it shows that 292 of the 438 “Refer” input class was correctly classified and 146 were not. Similarly the bottom row shows the summarized results for the “Do Not Refer” input class. One hundred thirty were correctly classified, and one hundred fifty five were incorrectly classified. So, the top-left and bottom-right cells are the correctly classified data, and the bottom-left and top-right are the false-negative and false-positive cases, respectively.

		J4.8 Classification	
Input Class		Refer	Do not refer
	Refer	292	146
	Do not refer	155	130

Figure 2.4: A confusion matrix

2.2.3 Sensitivity, Specificity, and Accuracy

Sensitivity, specificity and accuracy are statistical measures of binary classification tests and are analysis tools typically used by researchers in the medical community. In relation to the research of this paper, sensitivity is defined as the proportion of the number of cases with vision disorders that were correctly classified as such by the classifier. Specificity is defined as the proportion of cases without vision disorders that were correctly classified as such by the classifier. Accuracy is defined as the overall correctness of the classifier to the data presented. Ideally, the system would have a sensitivity of 100% and a specificity of 100%, but, in reality, there is a trade off between the two values. The value in these measurements is revealed when dealing with data that is skewed unevenly toward one input class. For example, if you started with a dataset consisting of 80% referral classes, then you could have a classifier perform with an 80% accuracy just by classifying every input class as refer. Obviously, this would not be a very good system, and that number does not accurately represent the performance of the classifier; the sensitivity would be 100% and the specificity would be 0%. Using the data from figure 2.4 for a specific example:

$$Sensitivity = \frac{292}{292+146} = 66.6\ percent$$

$$Specificity = \frac{130}{155+130} = 45.6\ percent$$

$$Accuracy = \frac{292+130}{292+246+155+130} = 51.3\ percent$$

2.2.4 K-Fold Cross-Validation Testing

K-fold cross-validation, sometimes called rotation estimation, is a classifier testing methodology that attempts to overcome some of the weaknesses described about the holdout testing methodology [Kohavi 1995]. In this case, the dataset is divided into K mutually exclusive data subsets of equal size. The classifier is then trained on K-1 of the datasets and tested with the remaining dataset. This train/test cycle is repeated with the classifier K times, and on each iteration, a different dataset is used a the test [Kohavi 1995]. Then the cross-validation estimation of accuracy is calculated by averaging the overall number of correct classifications divided by the number of those instances in the dataset [Kohavi 1995]. This method can be further improved by using a stratification of the original dataset. This means that the data is selectively stratified to have roughly the same proportions of the classes in each dataset as are represented in the overall dataset [Kohavi 1995]. The effect of stratification is that a particular dataset would not be abnormally represented in a particular K dataset [Kohavi 1995].

2.3 Amblyopic Risk Factor

The patients who are at risk of developing amblyopia typically exhibit certain factors. Beyond the socio-economic risk factors that sometimes lead to undiagnosed predisposing

conditions, there are distinct medical and genetic problems present the population with a greater risk. The socio-economic factors are similar to other medical issues that arise when a population does not have reliable access to medical care. This is often seen in poorer regions and underdeveloped countries, but the root problem is the same: the strabismus or anisometropia goes undiagnosed and leads to the development of amblyopia [Steinman 2000].

Since amblyopia is a developmental disorder, the medical conditions seen are ones exhibited by young children. Infants are four times as likely to develop amblyopia when they are premature, small for their gestation dates, or have a genetically related family member who could have developed amblyopia. In addition, children with neuro-developmental delay are six times more likely to develop the condition. The group at the highest risk are the children that are deprived of visual stimulation early in life, typically before age two. In all cases, the earlier in post-natal life the predisposing condition presents, the greater the impact on vision. In addition, the longer the condition persists without treatment, the more profound the level of amblyopia [Steinman 2000].

2.4 Artificial Intelligence Techniques

2.4.1 Decision Trees

Decision tree learning uses a decision tree as a predictive model and maps observations about an item to conclusions about the item's target value. In these tree structures the leaves represent classifications and branches represent the joining of features that lead to those

classifications. Decision tree learning is a common method used in data mining and is a model of the data that encodes the distribution of the class label in terms of the predictor attributes [Pao 1989].

The idea behind a decision tree is fairly straightforward. It is similar to how a person might make a decision to put on a coat, "Is it cold → yes, then put on a coat. → no, then do not put on a coat". In addition, you are able to induce a decision tree by example when you have a training set; thus more complex data is modeled more easily by presenting the training set and then building the tree based on that data set [Russell 1995]. The problem of finding a decision tree that agrees with the training set might seem difficult, but it is fairly straightforward. It would involve constructing a tree that has one path to a leaf for each example. Then, when given the same example again, it will output the correct classification [Russell 1995]. However, the problem with this trivial tree is that it just memorizes the observations. Since the goal of a classification algorithm is to work well on the general cases and to extrapolate to examples that were not part of the training set, a set of algorithms to prune the tree to a more generalized form are typically used [Russell 1995].

A common algorithm used to generate a decision tree named C4.5 was developed by Ross Quinlan in 1993. It uses the concept of information entropy from a set of training data to build the decision tree [Quinlan 1993]. The algorithm uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. C4.5 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is

the one used to make the decision. The algorithm then recurses on the smaller sublists [Pao 1989] [Quinlan 1993].

This project will use the J4.8 implementation of the C4.5 algorithm. This implementation was built by researchers are the University of Waikato in New Zealand (Weka) and is distributed for use in research projects such as this. According to the researchers, the J4.8 implementation is modeled after C4.5 revision 8, the last published algorithm by Quinlan. Quinlan has continued to improve on his data mining algorithms; however, he no longer publishes the details of the new algorithms. The J4.8 is essentially the same as C4.5 but has performance improvements for space and complexity during tree generation [Witten 2005].

2.4.2 Random Forests

Random forest classification is a supervised learning algorithm that builds on the decision tree algorithm. The original concept was first proposed by Tin Kam Ho of Bell Labs in 1995 where the random forest grows many classification trees. The values represented in each tree are a random vector sampled from the original dataset with the same distribution for all trees in the forest [Breiman 2001]. Once trained, the classifier takes a new input vector (pattern) and sends it down each of the trees in the forest. Each tree gives a classification, and the tree "votes" for that class. The forest chooses the classification having the most votes [Breiman 2001] [Kam Ho 1995].

Each tree is grown as follows:

1. A training subset is defined from the original data set, then a subset of the training set is sampled at random [Breiman 2001].

2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning [Breiman 2001] [Kam Ho 1995].

In general, the random forest classifier is the most effective tree algorithm for problems involving highly dimensional data because of the existence of more subspaces [Kam Ho 1995]. Key benefits of using a random forest algorithm over a decision tree include the random forest classifier's ability to negate the effects of over-fitting. In statistics and in classification problems, over-fitting occurs when the model that results from a supervised training algorithm describes random error or noise instead of the underlying pattern. A classifier that has been over-fit to a dataset will generally have poor predictive performance and will exaggerate the minor fluctuation in data.

2.4.3 Artificial Neural Networks

Artificial neural networks (ANN), also referred to as a multi-layer perceptron, is a technique that seeks to mimic the biological brain learning function. It is a mathematical or computation model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information from an input layer through one or many hidden layers and provides a response at the output layer [Pao 1989].

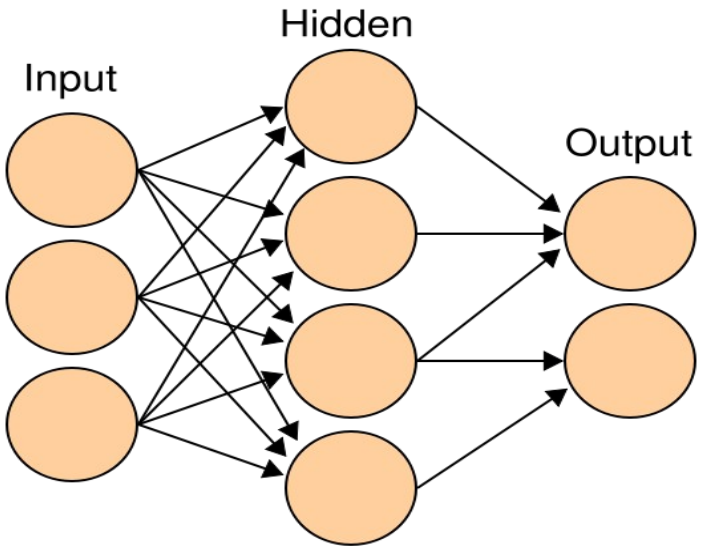
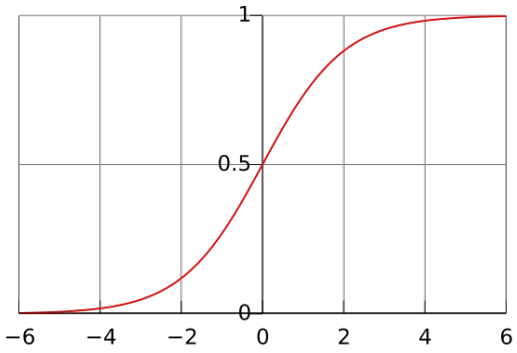


Figure 2.5: Multi-layer perceptron
Courtesy of Wikipedia Foundation Inc.

For this project the ANN will be used as an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. It is a feed forward network that will use a back propagation supervised learning algorithm during the training phase. A feed forward network means that an input vector will be presented to the network at a set of input nodes. The stimulus will travel through each layer of the network and will be transformed at each layer through a summation and an activation function. In the network used during phase one of this research paper the activation function is sigmoid [Pao 1989][Mathworks]. The primary purpose of the activation function is to amplify or attenuate the output based on a weighting factor that is adjusted at each node [Pao 1989]. Figure 2.6 is a graphical representation of the affect of the sigmoid activation function has on the data, further illustrating the way in which the values are constrained between zero and one.



$$o_j = \frac{1}{1 + e^{-(net_j + \Theta_j)/\Theta_o}}$$

Where net_j is the input from the previous layer, Θ_j serves as a threshold or bias, and Θ_o controls how sharply the function varies from zero to one [Pao 1989].

Figure 2.6: Sigmoid Activation Function

During the training phase of the classifier, the output of the network will measure the difference between the desired output and the actual output. The algorithm will then propagate weight changes back through each layer of nodes based on the difference between the desired and actual outputs. This is referred to as the generalized delta rule with back propagation of error [Pao 1989]. The training algorithm will continue to iterate over the training data sets until a desired mean squared error is reached between the desired output and actual output.

Chapter 3: Related Work

Three methods are currently used to identify amblyopia. The first two approaches are well known and commonly used: traditional vision screening and photorefractive screening [Kemper 2007]. The third method is an automated vision screening system that was developed recently and is still being perfected [Van Eenwyk 2008].

3.1 Traditional Vision Screening

Traditional vision screening is based on the identification of symbols. This is either through the Snellen E Chart or the Stycar Graded Balls [Kemper 2007]. This is probably the method that is most familiar to the public and involves the patient sitting with a trained specialist or ophthalmologist.

The traditional Snellen E Chart is printed with eleven lines of block letters, where the first line consists of one very large letter and the subsequent rows have increasing numbers of letters that decrease in size. A patient taking the test is located 20 feet from the chart, covers one eye, and reads aloud the letters of each row, with the smallest row that can be read accurately indicating the patient's visual acuity in that eye. For example, a patient who needs to stand 20 feet away from a target that could be seen at 40 feet by a standard patient is said to have "20/40" vision. This method of measurement is the source of phrases such as "20/20" and "20/40".Figure 2.1 shows an example of the Snellen E Chart.

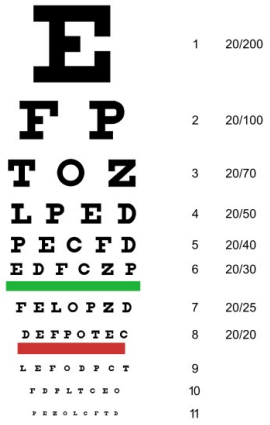


Figure 3.1: Snellen E Chart used in traditional vision screening
Courtesy of Wikipedia Foundation, Inc.

The Stycar Graded Ball test involves using a white ball before a dark background with a specialist observing the patient. The specialist will hold the white ball approximately 1.2 feet from the patient and measure the time needed for a patient to fixate on the ball. In addition, measurements of peripheral vision are also taken using this method [Sheridan 1973].

The ophthalmologist then uses these two well-known tests to identify if the patient is at risk for developing amblyopia. In the case of this research project, the traditional vision screening was conducted personally by Dr. Cibis and the technicians at his clinic. He was very thorough in his screening of these patients in an effort to make the data the "gold standard" by which the software can both be trained and measured against [Kemper 2007]. Notably, the human factor is another reason that this sort of an automated screening system is desired. Not only does an expert software system provide consistency to a preliminary diagnosis, it also can overcome human deficiencies and the plain fact that some medical doctors are better

at screening than others. When an expert software system makes the initial screen, the medical doctor can focus his or her attention on the patients who are at risk.

3.2 Photorefractive Screening

Photorefractive screening is based on a system to interpret the images of the eyes. It will not directly identify amblyopia, but it will look for defects in the eyes that may lead to amblyopia. Many photorefractive screening systems are in development, and some are commercially available. Van Eenwyk evaluated four of the more advanced systems in order to compare their published results with the results of the Automated Video Vision Development Assessment (AVVDA) [Van Eenwyk 2008].

The Photoscreener system is based on single images of the patient's eyes. The system is a hand held camera with instructions for the operator on how to identify potential problems. The operator will the take a picture of each eye and then analyze the resulting frame per the instructions. If the frame shows markers of a vision disorder, the patient is identified for referral [Van Eenwyk 2008].

The RetinoMax K-Plus 2 and the SureSight Vision Screener are two automated screening systems that are commercially available. The benefit of this sort of system is that the subjective nature of an operator's analysis is removed from the equation, and a more consistent result should be evident [Van Eenwyk 2008]. However, the drawback to this sort of system is that it operates on single frame analysis and not on a video of the patient looking

at a light source [Van Eenwyk 2008]. In addition, it is more focused on identifying general vision problems than those related to amblyopia [Van Eenwyk 2008].

The final product reviewed by Van Eenwyk was the Pediatric Vision Screener. Instead of taking pictures with a recording device, this screener measures the frequency of the polarized light off the retina as a light source circles the eye [Van Eenwyk 2008]. By comparing the results of the test on both eyes, the system is able to identify strabismus [Van Eenwyk 2008]. According to Van Eenwyk, the results are reasonably accurate, but the system requires the cooperation of the patient, a problem when dealing with children younger than six [Van Eenwyk 2008].

MTI Photoscreener and Visiscreen 100 are two additional photorefractive screeners identified during this research [Freedman 1992]. They are described here just as examples of other commercially available screening systems but are not used in comparison to AVVDA. According to the researchers, both photo-screening systems are not automated and require the systems to be operated and analyzed by skilled professionals [Freedman 1992].

While all of these systems hold promise, they all fall short in one of three ways. The first is they require patient cooperation that would typically be beyond the demographic that would be helped by the amblyopic screening. The second is they operate only on single frame analysis and will miss the subtle foveation that Dr. Cibis theorizes will more accurately identify a problem. Finally, they suffer from the same issue as Dr. Cibis VVDA system: A

well trained operator or a medical doctor is required to analyze the data for the factors indicating a problem.

3.3 Automated Photorefractive Screening

The final method is an automated photorefractive screening system. This type of system differs from the previously mentioned systems at the analysis step. The system works by having an operator take a short video of the patient, which is then analyzed by the software in the following manner:

- First, the software identifies the frames where both eyes are open and looking at the light source. These frames are identified as key frames.
- The software will isolates the location of the eyes and pupils in the key frames.
- Finally, the software uses various techniques to extract the distinguishing features that may be indicators for amblyopia.

At this point, various machine learning techniques are utilized to produce a binary output: either recommend a referral to a specialist or not [Cibis 2005] [Van Eenwyk 2008]. This project will expand upon the automated photorefractive screening method. The primary work will be to investigate a different set of features and compare the results against the current methods.

3.4 Video Vision Development Assessment

In a thesis written by Dr. Gerhard Cibis in 1994, a method of identifying strabismus and amblyopia using video images was researched and presented [Cibis 2005]. The method involves a low-cost, consumer grade video camera with a light source attached to the base of the camera. The patient would sit approximately 52 inches from the camera and look at the light source while approximately two minutes of video is recorded. Dr. Cibis named this system the Video Vision Development Assessment (VVDA). Figure 4.1 shows a picture of the device.

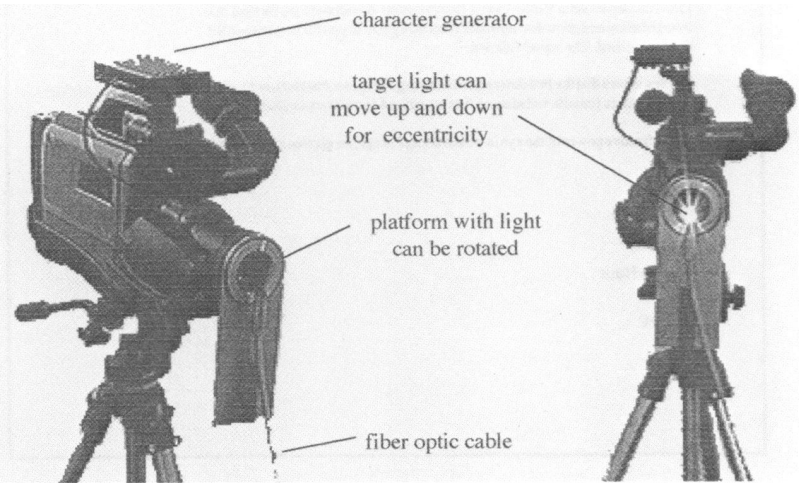


Figure 3.2: VVDS System [Cibis 2005]

The video would then be shipped to a lab for digitization and analysis by medical doctors or trained technicians. The processing of the video at the lab extracts the features that the technicians will use to determine if the patient should be referred to a specialist. This processing is generally divided into frame selection and feature extraction.

Frames are selected using the following criteria:

1. The patient's head is in the primary position.
2. The frame is in focus, with the Hirschberg point at its smallest.
3. At least one pupil should be at its largest diameter.
4. The Hirschberg point should be just nasal to the center of the pupil.

The frames selected are then measured, and the following features are extracted:

1. Pupil width is measured in pixels.
2. The distance between the center of the Bruchner reflex and the medial edge is measured.
3. Any observable crescent is measured in length and position.
4. Luminance of each pupil is measured with a circular histogram.
5. The resulting linear measurements are converted from pixels to millimeters and analyzed by the system.

Finally, once the data is extracted and all the frames processed, the technician is able to make a refer/no refer decision for the patient. While this process begins the steps of creating an automated system, it still suffers the same issue as the traditional methods: it requires a trained technician or M.D. to analyze the data. In addition, it adds a delay to the processing since the recorded data would need to be sent to a processing center before analysis can begin.

3.5 Bruchner's Reflex

One of the key features that Dr. Cibis VVDA system reviewed were features derived from the Bruchner's red fundus reflex test [Cibis 1994]. Based on the system perfected by Bruchner in the early 1980s, VVDA reviews the photographs of the key frames for a pass or fail of the Bruchner test [Cibis 1994]. The results of the Bruchner test are the first decision in the decision tree algorithm and have the largest weight associated with the results of that test. The physical attributes reviewed are not focused on a certain value that the reflex test be equal in order to pass; on the contrary, the real measurement is to review the difference between the two eyes and render a verdict based on the size of the variance between the two eyes [Cibis 1994]. The brighter eye will be the one that is deviated due to the way the light "leaks" into the observer's eye. In a normal eye, the reflection would shine back at the light source on which the patient is fixating [Cibis 1994].

3.6 Automated Video Vision Development Assessment

More recently, Dr. Cibis collaborated with computer science researchers from the University of Kansas and the University of Missouri in an attempt to automate the analysis of the frames using artificial intelligence techniques [Wang 2005][Wang 2002][Van Eenwyk 2008].

One of those researchers was Wang, and his primary focus was to implement the image processing and case based reasoning algorithms that constituted the first version of the Automated Video Vision Development Assessment (AVVDA). The details of how the images were processed is covered in two papers published Wang in 2002 and 2005 as a part of his dissertation requirements at the University of Missouri. In a completely automated

fashion, he is able to identify the key frames, isolate the pupils, and locate the Hirschberg point [Wang 2005][Wang 2002]. So, based on the work of Dr. Cibis, Wang implements the video processing and feature extraction pieces in this system. Figure 4.3 shows the image output from video processing system. The Hirschberg reflex and iris diameter are highlighted.

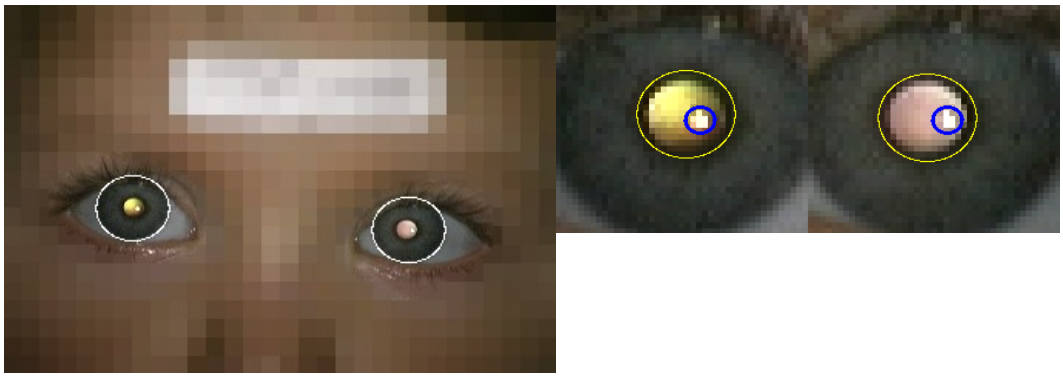


Figure 3.3: Key frame output from the AVVDA system [Van Eenwyk 2008]

The second researcher with whom Dr. Cibis collaborated with was Van Eenwyk. As part of a masters thesis, Van Eenwyk made use of the feature set collected by Wang but with a different set of classifiers. He measured the efficacy of using a different set of classifiers in order to come up with a refer/no refer decision within the software. The overall goal of the AVVDA system is to allow an unskilled technician to operate the system and accurately get a decision about patient referral to an optometrist or ophthalmologist [Van Eenwyk 2008].

In the current form, AVVDA uses cased based reasoning, C4.5 decision tree, and artificial neural network classifiers to assist in making the decision. These three classifiers are further explained in chapter three. Training the classifiers involves 54 features that are extracted from the images. The desired output uses the ophthalmologist official recommendation for the patient, refer (1) or do not refer (0). Finally, the testing of the system uses 10-fold cross validation testing methodology to ensure that the results are accurate. As of the publication of the Van Eenwyk paper in 2006, 610 patient videos had been collected for processing using the AVVDA system. As a part of the collaboration between Dr. Cibis and Van Eenwyk, Dr. Cibis was responsible for collecting the videos. The videos were collected when he would visit schools for standard screening appointments or when patients would see him at his optometrist offices.

Notably, additional work was done by Van Eenwyk that involved chaining the classifiers in an attempt to boost the accuracy [Van Eenwyk 2008]. However, since this paper will focus on using single classification methods, those results have been omitted from the proposed methods. This is in an effort to compare the results of Van Eenwyk with a new set of features extracted using similar single classifiers. Future work is reserved for use of chaining the classifiers. Figure 4.4 shows the accuracy results of the primary classifiers tested in the research [Van Eenwyk 2008].

Classifier	**Sensitivity**	**Specificity**	**Accuracy**
Case-Based	84.60%	58.60%	75.20%
ANN	61.50%	63.60%	62.30%
C4.5	76.20%	79.50%	77.40%

Table 3.1 Accuracy results of the AVVDA system [Van Eenwyk 2008]

Chapter 4: Methodology

For this project we propose the use of an additional group of features based on the color slope that is extracted from the key frames. So this project is additional research for the aforementioned AVVDA system and will build on the work of Dr. Cibis, Wang, and Van Eenwyk [Van Eenwyk 2008].

The new group of features will then be input into three classifiers and the accuracy of the classifiers will be compared. The three classifiers to be used are a C4.5 decision tree classifier, a random forest classifier, and an artificial neural network classifier [Quinlan 1993] [Mathworks] [Witten 2005]. The Weka software implementation will be used for the C4.5 decision tree and random forest [Witten 2005]. Matlab will be used for the artificial neural network [Mathworks].

4.1 Experiment Setup

Experiment setup involves processing the patient videos and prepping the data for feature extraction. Since the work of Van Eenwyk, the available patient data has grown from 610 to 723 patient videos.

Noting the quality of the data that has been obtained is important. Most of these videos are from visits by Dr. Cibis to local or regional primary schools; therefore, they should not be skewed toward patients who were referred by an ophthalmologist. Since the amblyopic condition statistically affects two to five percent of the population, the data is predicted to

show the majority of the patients to be healthy. However, further analysis of the data shows the opposite, that there are a majority of referral decisions (457) over non-referral decisions (288). This could be because the physician examining the patient is being overly cautious; however, more likely is that the voluntary nature of the experiment causes the parents or guardians to encourage the child in their care to see the doctor. Since that encouragement typically only happens if the parent or guardian suspects a vision problem, the data would be skewed toward patients with vision problems.

Experimentation was at the core of this research project. The overall goal was to find a better set of features that could be extracted from the patient images and submitted to the artificial intelligence classifiers for a more accurate result than previous work provided. For that reason, feature extraction work and efforts to reduce the number of features in the dataset (dimension) is at the center of the project and encompassed the majority of the time. The research is broken into three phases, with the primary differences between the phases reflected in how the features are extracted. The next sections will describe the work done in each of the phases in greater detail. The majority of the data manipulation work is similar for all three phases and at the steps where the phases differ, those differences will be discussed.

4.1.1 Feature Extraction Overview

The overall goal of the feature extraction step is to represent the rate of change of the colors from one side of the pupil to the other. In order to get this color data in the affected area of the eye, the raw image data must go through a series of steps to extract the information that will be submitted to the classifiers. Figure 4.1 diagrams the general procedure.

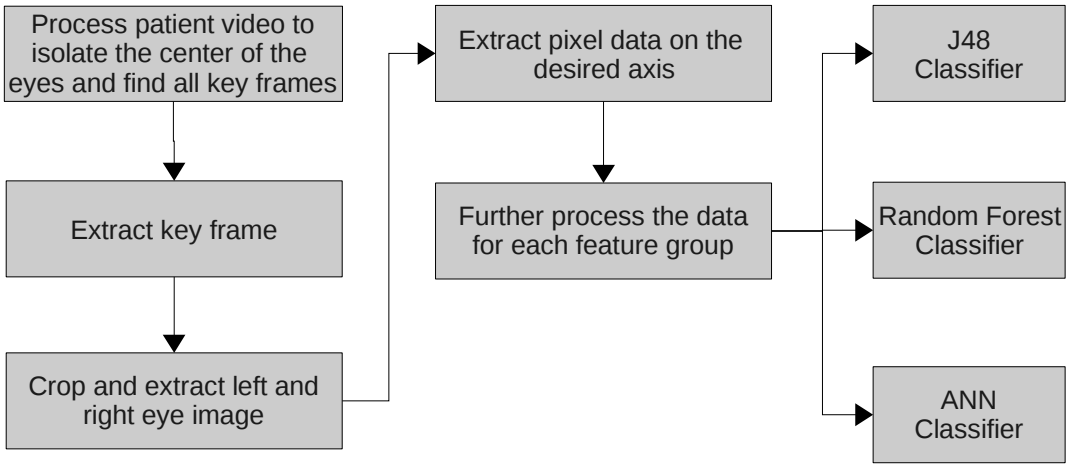


Figure 4.1: Diagram of the feature extraction process

4.1.2 Processing Patient Videos

Processing patient videos takes the footage of the patient, processes the frames for the key features, and identifies the key frames. This is the work that was done by Wang and described previously in this paper.

4.1.3 Extract Key Frame

The next step involves selecting a key frame from which to extract the pixel data from. This is an important step because the hope is to use a frame in which the patient is exhibiting the amblyopic condition. As explained in a previous section the key frame identification process primarily focuses on ensuring that the patient's head is in a primary position, the image is in focus, the Hirschberg point is at its smallest, at least one pupil is at its largest diameter, and the Hirschberg point is just nasal to the center of the pupil.

During phase one and phase two, the research takes a very simple approach to identifying the one frame that will be used for feature extraction. It will take the frame in the exact middle of the stack of frames. If there is an odd number of frames it will take frame that is whole number division of the entire frame count by two. The goal is to avoid the fringe frames (the first or last), and analysis of a subset of the data shows that the middle frame has the greatest chance of being one of the best images.

During phase three, the frame selection process was altered, and instead of selecting the middle frame, a different process determined the frame. The goal for phase three is to select the frame out of all the key frames where the patient's head is most level. Since the entire classification problem is centered on the reflection of light off the lens and the retina, keeping the angle of reflection the same for both eyes would, theoretically, further enhance any differences between a healthy eye and the eye of a patient who should be referred to a specialist. This idea is further supported by the important features extracted around the crescent reflection and the Hirschberg point that is measured from that crescent. The frame

that is most level is determined by the pixel location of the center of the pupil at the y axis. The value is compared between the two eyes for each key frame and the one where the value is the closest is the frame selected. When no key frame could be found where the difference in the height of the eyes was less than five pixels, then the patient data was discarded. The data was discarded so that the classifiers would not be trained with imperfect data and in hopes of further isolating the features that will give the best results. This process reduced the number of patients in the entire sample from 723 to 499.

4.1.4 Crop and Extract Eye Images

Once the key frame is identified, the image processing has determined the x and y coordinate for the center of the pupil and the radius of the iris. Using this data a left and right eye image is produced. Both images are 62 by 56 pixel images and resemble figure 4.2.

Figure 4.2: An example of a key frame cropped around the pupils for each eye

4.1.5 Extract Pixel Data

After the individual eye images are available and cropped, the pixel values are to be extracted from different angles off the center of the eye. During phase one of the project these angles are the 0, 45, and 135 degrees on the center of the eye and across the iris and the pupil. The 90 degree axis will not be used because it typically is partially obstructed by the upper or lower eyelid. Therefore, on the zero axis, 36 pixels are extracted, and on the 45/135 axes, 28 pixels are extracted. That makes a total of 92 pixels times three colors on each pixel (red, blue, and green). Finally, that number is doubled to account for both the left and right eye. Notably, each pixel extracted was averaged against its immediate neighbor in order to smooth out any drastic errors that may have occurred in the image capture process. Figure 4.3 shows where the data is extracted across the eye image.

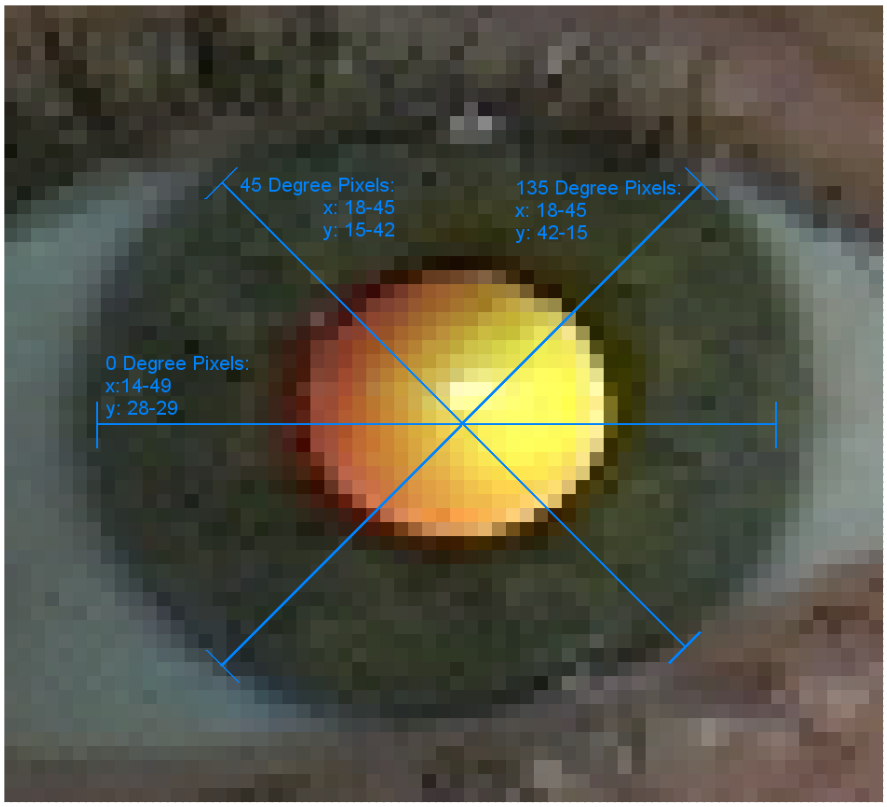


Figure 4.3: Pixel color extraction

Phase two and three used a different process based on conversations with Dr. Cibis after reviewing the results of the first phase. Since iris color had been evaluated by Van Eenwyk in his thesis and did not show any predictive capability for strabismus and anisometropia [Van Eenwyk 2008], the decision was made to isolate the color extraction to the pixels in the pupil. Also, in order to reduce the dimensionality of the data and focus on the areas that Dr. Cibis theorizes hold the most predictive information, the axes extracted were focused on the vertical center of each eye (90 degree) and the axis that directed through the nose (45 degree for the left eye, 135 degree for the right, both from the observer's point of view). This

process reduced the number of raw pixel features to 180. Figure 4.4 diagrams the phase two and three feature extraction.

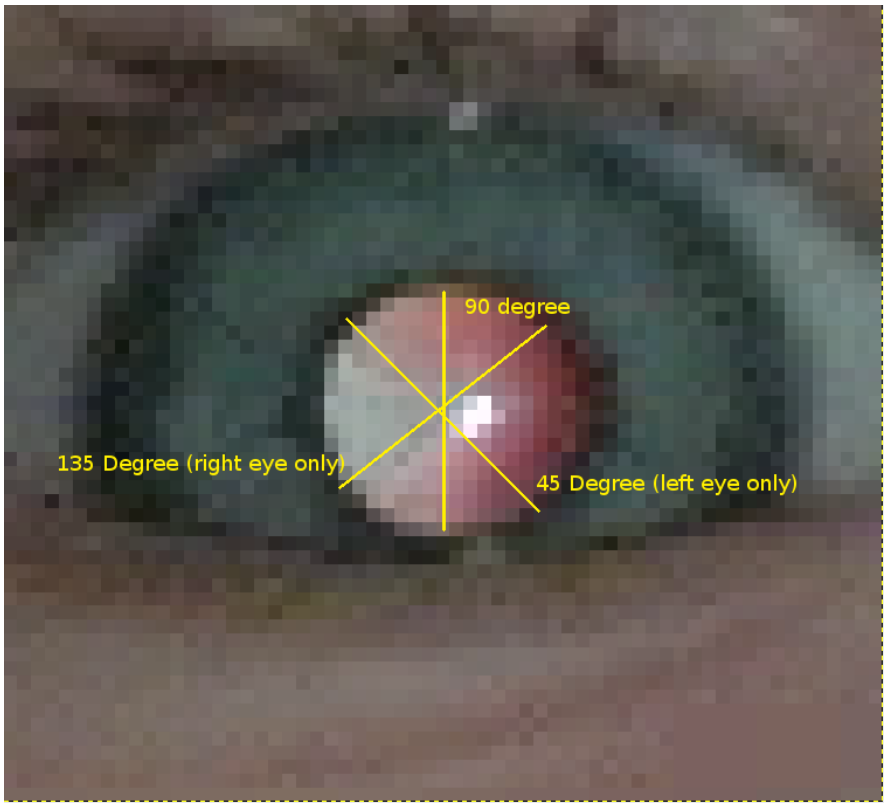


Figure 4.4: Phase two and three pixel color extraction

4.1.6 Produce Feature Groups

Since the goal of this project is to investigate feature groups to determine if they will provide enough distinction for accurate classification, four feature extraction strategies have been chosen. Not all the strategies are used in all three phases. Phase one uses all four as an initial review of how the data might perform. Phases two and three only use the Pixel Values feature group and the One Color Slope Values feature group. The reason that the PCA Pixel

Values and the Four Color Slope Values feature groups are not used in the later phases is because they are both dimension reduction strategies. Since the phase two and three datasets use significantly fewer data points, reducing the dimensions of the raw data was not seen as necessary.

Pixel Values

Here the strategy is to use all the pixel color values that were extracted from the eye images. While it is not expected to yield good results, it is a starting point for establishing the efficacy of the data. Assembly of the data is fairly straightforward, and input files for the classifiers are produced by appending the Red, Blue, and Green (RGB) values for each axis and each eye.

PCA Pixel Values

The raw pixel values produced from phase one will probably be prone to over-fitting of the data in the classifiers, since the 552 features almost match the number of patients in the dataset. Therefore, the values were processed with principle components analysis (PCA) to reduce the dimensionality of the data. The goal is to find a balance between a small number of features and a high aggregate participation in the overall variance of the data.

The data was processed in the following steps:

1. Start with the 552 features and standardize the data. When using PCA, a good practice is to standardize the data in order to find the true dimension variance. A common standardization formula was used:

$$\frac{(X_i - \mu)}{\sigma}$$

where X_i is the vector X minus the mean μ and divided by the standard deviation σ of the vector X.

2. Calculate the covariance matrix using the standardized dataset.
3. Run PCA against the dataset.
4. Transform the data and compute the principle components.
5. Reduce the dimensions based on the percentage of participation in the variance.

After the data was processed, 15 principle components were selected. These accounted for 74.23% of the participation in the total variance of the dataset.

Four Color Slope Values

In this phase one feature group, the rate of change in color is extracted from four segments of the color line for each axis. The formula for calculating the slope of each segment allows the rate of change between every data point to contribute to the final slope value.

$$\frac{[\sum_{i=1}^{j} (p_i + p_{i+1})]/(j-1)}{j}$$

In the formula, j is the number of pixels in the segment, and p is the pixel color for that particular color band (red, green, or blue). Once processed, 64 features result: 4 slope values x 3 colors x 3 axes x 2 eyes. Figure 4.5 is an example color line with the segments diagrammed.

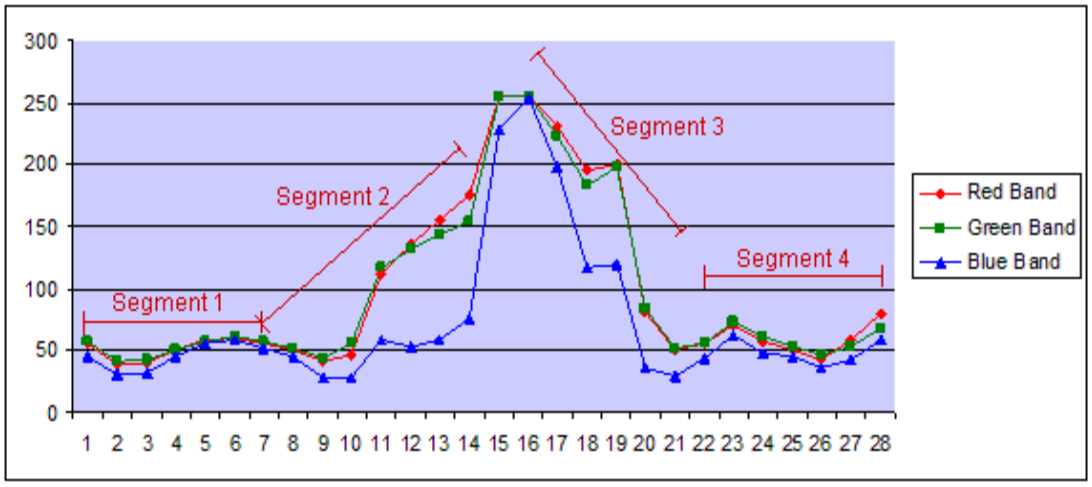


Figure 4.5: An example of the color segments along the 45 degree axis

One Color Slope Values

The final feature group is drawn from taking the slope across the entire color line. This is similar to the four color slope values in that it will use the same formula for calculating the slope (formula 3.5). The difference is that the "j" variable will be equal to the length of the entire color line. Once the data is extracted and calculated, 18 features are produced in phase one (3 colors x 3 axes x 2 eyes) and 12 features are produced in phase two and three (3 colors x 2 axes x 2 eyes).

4.2 Classifier Training

The classifiers to be used have multiple parameters that will affect their behavior and classification accuracy. The J4.8 Decision Tree and Random Forest classifiers were trained with data from all phases, while the ANN was trained with data from phase one. After reviewing the performance of the ANN with respect to the other classifiers, the determination

was made to focus the remaining research efforts on the decision tree classifiers. The ANN did not perform much better than flipping a coin.

4.2.1 J4.8 Decision Tree

The Weka J4.8 decision tree classifier allows for some customization of its operation based on command line options. These options will tune the performance of the classifier and could possibly lead to better classification results [Witten 2005]. The Weka J4.8 documentation exposes ten tunable options; however, the most pertinent are the confidence threshold and the minimum number of instances per leaf. The confidence threshold allows the user to customize the pruning of the tree. Decreasing the value of the parameter has the effect of more aggressively pruning the tree and removing leaves that do not substantially contribute to classification. The minimum number of instances per leaf will affect how the tree is constructed and cause further splits in order to meet the minimum number of classes at the leaf layer [Witten 2005].

For all of the J4.8 classifier training in this project the default parameters were used. This means that a 10-fold cross validation is used for testing. Pruning is used with a .25 confidence threshold and a minimum of two instances per leaf [Witten 2005]. Further experimentation was done with the other options, however, the results were not any better than what was produced by the options discussed. For an example of a pruned decision tree, figure 4.6 shows the tree produced from the phase two, one-slope feature set.

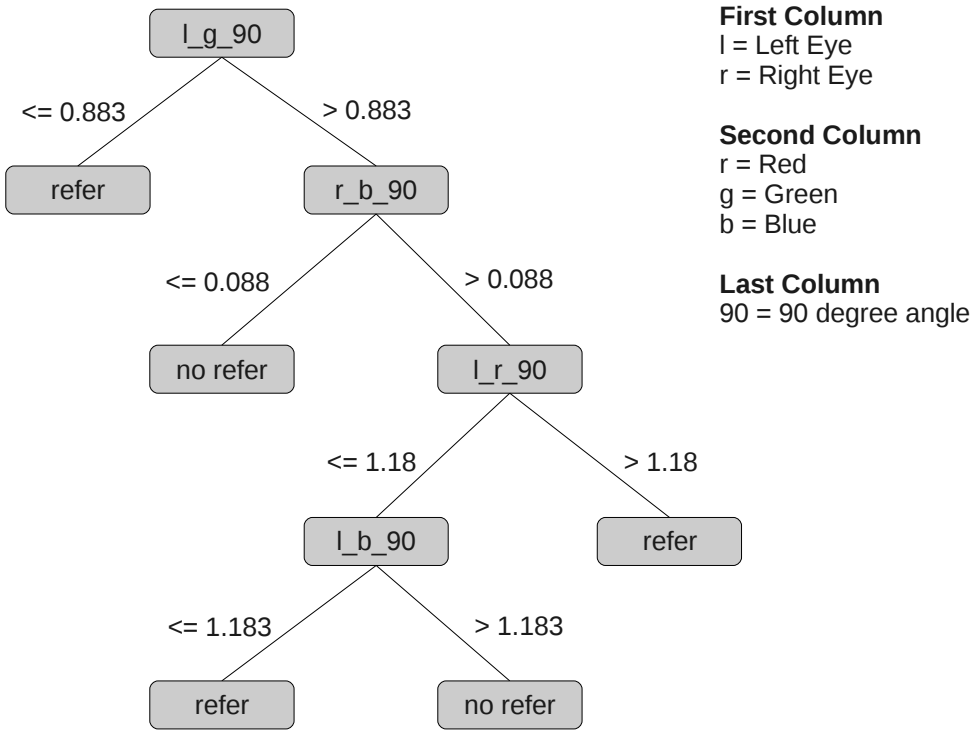


Figure 4.6: Decision tree generated from the phase two one slope feature set

4.2.2 Random Forest

The Weka random forest classifier allows for customization of its operation based on several command line options: the number of trees produced, the number of features used, and the depth of each tree. The random forest produced the best results by building large numbers of trees, typically greater than 1000, using all of the features, and allowing the depth of the trees to grow as large as necessary.

When testing with the data from each of the phases, the typical behavior observed would be an increased accuracy rate the larger the number of trees used. However, at a point in the train and test cycle, the accuracy would plateau. At this point the train and test cycle were halted, and the best accuracy was deemed to have been found. Brieman further supports this behavior through mathematical examination and explaining that random forests do not over-fit as more trees are added, but produce a limit on the generalized error [Breiman 2001].

4.2.3 ANN

When using the artificial neural network implementation provided by Matlab, the default settings were taken when the "newff" function was called [Mathworks]. In general, the inputs were not normalized between 0 and 1, but were kept at their original values, 0 to 255. The decision was made to leave them in the raw form because there was already a reasonable upper and lower bound on the data and there would not be any fluctuations or spikes that would cause the classifier to miss any subtleties in the data due to an overpowering spike. The data spike would be similar to a person attempting to identify a sound however part of the sound is extremely loud but the portion that allows accurate identification is very quiet.

The exception to the normalization process is in the case of the principle components analysis in phase one. The reason for normalizing the dataset at that phase is due to the process of gathering the principle components. Part of the procedure is to start with normalized data. The procedure used was discussed in section 3.2.6. Finally, the activation function of the ANN, also called the transfer function, used with newff is the sigmoid activation function at the hidden layers with a purely linear activation function at the output layer.

In order to find the best combination of hidden layers and number of nodes per hidden layer, a Matlab function was written that will systematically test three hidden layers with up to 20 nodes in each layer down to one hidden layer with five nodes. On each iteration and test cycle the program would attempt to converge the network on the training set at a Mean Squared Error (MSE) of 10e-10. Once the network converged or not a significant shift in the MSE caused the training process to exit, the network was tested against 10% of the data (72 patients). The program would then log the results of the test and then decrement the number of nodes in the layer.

Importantly, the training set and testing set were divided so that the network did not train with any data from the test set. The test set was randomly selected from the entire population using the Matlab "rand" function. This means that the ANN did not use the 10-fold cross validation testing methodology. This decision was made due to the amount of time needed for the computer to train and test the networks and the poor performance of the classifier after the first validation tests.

Chapter 5: Experimental

5.1 Phase One

The tables 5.1 through 5.4 show the confusion matrix results of the phase one classification results. Tables 5.5 through 5.8 summarizes the performance of each classifier with the data sat using specificity, sensitivity, and accuracy measurements.

5.1.1 Raw Pixel

		J4.8 Classification	
Input Class		Refer	Do not refer
	Refer	292	146
	Do not refer	155	130

		Random Forest Classification	
Input Class		Refer	Do not refer
	Refer	361	77
	Do not refer	155	130

		ANN (3 Hidden layers, [5 4 2]) Classification	
Input Class		Refer	Do not refer
	Refer	23	20
	Do not refer	14	15

Table 5.1: Pixel value result detail

5.1.2 PCA Pixel

		J4.8 Classification	
Input Class		Refer	Do not refer
	Refer	411	27
	Do not refer	248	37

		Random Forest Classification	
Input Class		Refer	Do not refer
	Refer	374	64
	Do not refer	194	91

		ANN (1 Hidden layers, [Wang 2002]) Classification	
Input Class		Refer	Do not refer
	Refer	48	4
	Do not refer	20	0

Table 5.2: PCA Pixel value result detail

5.1.3 Four Color Slope

		J4.8 Classification	
Input Class		Refer	Do not refer
	Refer	258	180
	Do not refer	149	136

		Random Forest Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**353**	**85**
	Do not refer	**162**	**123**

Table 5.3: Four color slope value result detail

5.1.4 One Color Slope

		J4.8 Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**438**	**0**
	Do not refer	**285**	**0**

		Random Forest Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**340**	**98**
	Do not refer	**233**	**52**

Table 5.4: One color slope value result detail

5.1.5 Summary Results

552 Features	Sensitivity	Specificity	Accuracy
J4.8	66.67%	45.61%	58.37%
Random Forest	82.42%	45.61%	67.91%
ANN	53.49%	51.72%	52.78%

Table 5.5: Pixel value result summary

PCA (15 features)	Sensitivity	Specificity	Accuracy
J4.8	93.84%	12.98%	61.96%
Random Forest	85.34%	31.93%	64.32%
ANN	92.31%	0%	66.67%

Table 5.6: PCA Pixel value result summary

4 Slope (64 features)	Sensitivity	Specificity	Accuracy
J4.8	58.90%	47.72%	54.50%
Random Forest	80.59%	43.16%	65.81%

Table 5.7: Four color slope value result summary

1 Slope (18 features)	Sensitivity	Specificity	Accuracy
J4.8	100.00%	0.00%	60.58%
Random Forest	77.63%	18.25%	54.21%

Table 5.8: One color slope value result summary

5.2 Phase Two

The tables 5.9 and 5.10 show the confusion matrix results of the phase one classification results. Tables 5.11 and 5.12 summarizes the performance of each classifier with the data sat using specificity, sensitivity, and accuracy measurements.

5.2.1 Raw Pixel

		J4.8 Classification	
Input Class		Refer	Do not refer
	Refer	306	132
	Do not refer	185	101

		Random Forest Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**370**	**68**
	Do not refer	**194**	**92**

Table 5.9: Pixel value result detail

5.2.2 One Color Slope

		J4.8 Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**380**	**58**
	Do not refer	**224**	**62**

		Random Forest Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**438**	**0**
	Do not refer	**52**	**234**

Table 5.10: One color slope value result detail

5.2.3 Summary Results

Pixel (180 features)	Sensitivity	Specificity	Accuracy
J48	69.86%	35.31%	56.21%
Random Forest	84.47%	32.17%	63.81%

Table 5.11: Pixel value result summary

1 Slope (12 features)	Sensitivity	Specificity	Accuracy
J48	86.76%	21.68%	61.05%
Random Forest	82.19%	33.22%	62.85%

Table 5.12: One color slope value result summary

5.3 Phase Three

The tables 5.13 and 5.14 show the confusion matrix results of the phase one classification results. Tables 5.15 and 5.16 summarizes the performance of each classifier with the data sat using specificity, sensitivity, and accuracy measurements.

5.3.1 Raw Pixel

		J4.8 Classification	
Input Class		Refer	Do not refer
	Refer	86	107
	Do not refer	100	206

		Random Forest Classification	
Input Class		Refer	Do not refer
	Refer	72	121
	Do not refer	39	267

Table 5.13: Pixel value result detail

5.3.2 One Color Slope

		J4.8 Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**69**	**124**
	Do not refer	**78**	**228**

		Random Forest Classification	
Input Class		**Refer**	**Do not refer**
	Refer	**65**	**128**
	Do not refer	**59**	**247**

Table 5.14: One color slope value result detail

5.3.3 Summary Results

Pixel (180 features)	Sensitivity	Specificity	Accuracy
J48	67.32%	44.56%	58.52%
Random Forest	87.25%	37.30%	67.79%

Table 5.15: Pixel value result summary

1 Slope (12 features)	Sensitivity	Specificity	Accuracy
J48	74.51%	35.75%	59.51%
Random Forest	80.72%	33.68%	62.52%

Table 5.16: One color slope value result summary

5.4 Analysis

Notably, the ANN was tested using fewer samples because the 10-fold cross validation testing was not used. Instead, a holdout test was run with 10% of the population randomly selected as a member of the test set. However, the results showed as just slightly better than flipping a coin. Since the research was looking for a better feature set that the existing system, this sort of result is not good enough to warrant stratified cross validation or continuing to use the classification on the remaining phases of the project.

With the best overall accuracy of the new feature set at 68%, the results did not perform better than 77.5% accuracy achieved in the previous work done by Van Eenwyk [Van Eenwyk 2008]. The best accuracy was from the random forest, but that was almost 10% less accurate than the previous results. The large number of features may not have provided a distinct enough delineation between the two classes. In addition, the number of features was almost the same number of samples, therefore contributing to an over-fitting situation. The random forest likely produced the best results because it is able to overcome the over-fitting issue. The number of features and the number of samples lent itself to that sort of a problem. Figure 6.17 shows the comparison between the results of the AVVDA system and the results from this research project.

AVVDA (Previous research)			
Classifier	**Sensitivity**	**Specificity**	**Accuracy**
Case-Based	84.60%	58.60%	75.20%
ANN	61.50%	63.60%	62.30%
C4.5	76.20%	79.50%	77.40%

Current Research			
Classifier	**Sensitivity**	**Specificity**	**Accuracy**
Random Forest	82.42%	45.61%	67.91%
ANN	53.49%	51.72%	52.78%
J4.8	66.67%	45.61%	58.37%

Table 6.17 Accuracy results of the AVVDA system and current research

Chapter 6: Conclusion

6.1 Contributions

The research evaluated in this paper contributed to the advancement of the AVVDA system at the feature extraction step. While Dr. Cibis theorized that the color slope in the pupil of the patient eye would have been a good indicator of amblyopic conditions, the results show it to be less accurate than the existing system. So while that particular feature set proved to be less than satisfactory, this paper provided a thorough examination of color slope features paired with three distinct classifiers. At the testing step, the research also used a rigorous 10-fold stratified cross validation testing procedure for the two most promising classifiers, Random Forest and J4.8. This means that all patient data was taken into account and no attempts were made to filter out the difficult patient videos in order to gain a better classification accuracy.

6.2 Limitations

Two areas may have provided some limitations to the research. The first is the number of patient videos. As with any classification problem, the more high quality example data evaluated, the better chance of getting a good classification rate. In addition, the earlier videos taken were not as rigorously controlled as later videos, meaning that as the operators became more familiar with collecting patient data, they were also more detailed about distance and placement of the patient, both key to getting consistent data in which to train

with. Potentially, removing the earlier patient videos from the sample could improve classification results.

The second limitation may have been the concept behind the rate of change of color in the eye acting as a predictive feature. The concept is outside of Dr. Cibis' original thesis and is not reflected in any traditional screening test. As an example, the core elements of the original VVDA system involved analysis of the Hirschberg reflex and ratio, in conjunction with Bruchner reflex, crescent, and color saturation. Possibly, attempting to expand on those measurements may have yielded better results.

6.3 Future Work

Even though this project resulted in less than satisfactory classification results, the problem still has some promising research potential. During the work on this project, other methods became apparent and may yield better results.

6.3.1 Key Frame Selection

Finding the frame that captures the patient exhibiting signs of the amblyopic condition is one of the most important parts of the feature extraction process. It is also one of the most difficult. Theoretically, the signs might show in one frame or across multiple. One research area would be to investigate using more than one frame or an average of all the frames in hopes of capturing the necessary data.

6.3.2 Investigation of Foveating Frames

One complication discussed by Dr. Cibis is a phenomenon called microtropia or monofixation syndrome. The key feature is an intermittent deviation of six degrees or less from the true fixation point. Dr. Cibis chose the term “foveation” to refer to the true fixation to distinguish it from slightly off-axis fixation [Cibis 2005][Van Eenwyk 2008]. An examination of recorded video reveals frequent cases where the patient's eyes will switch between fixation and slightly off-axis fixation, or foveate. These patients will then switch again only a few frames later. The goal is to focus the feature extraction only on these foveating frames to determine if the color slope gathered from that frame will yield more accurate results than other frame selection strategies.

6.3.3 Include Color Slope with Previous Features

During the course of the research for this project, the new features extracted from the iris and pupil were not added to the existing features extracted by Wang and Van Eenwyk. The goal here would be to summarize the color slope features into a set of discriminating values and include them with the existing 54 features identified previously. Since the color slope in its raw form would substantially increase the dimensionality of the classification problem without a clear linear separability, a first step would be to distill the color features into a small enough feature set. PCA or some other sort of dimension reduction strategy could bring them under control. Then researchers could further test the existing system with the new feature sets and measure if the sensitivity, specificity, and accuracy numbers are improved as a result of including the color features.

6.3.4 Additional Feature Investigation

During the investigation of color slope, the primary focus was to work with the color data along specific axes. However, the entire pupil is illuminated as the patient is focusing on the light source. Since all of the information that allows a trained physician to make a decision is contained in that reflected light, some clue should be in the color map of the pupil. These clues will allow an expert system to make a decision based on the same visual cues that the physician would use. The research into this area would continue to analyze the reflected light and attempt to determine features that encompass the entire reflected pupil across the two eyes in the same frame. Two specific features were noticed during the color slope research that could hold clues.

The first is the color intensity measurements across both eyes. When the patient is identified as a referral, the intensity seemed to be out of sync between the two eyes.

The second is the focal point difference. The reflected light off of the back of the retina typically will provide a two pixel by two pixel focal point. The point is identified because it is the brightest spot in the pupil. For a healthy patient, this focal point typically shows at the center or inside of the center of the eye. In contrast, patients who should be referred seem to have the focal point to the outside on one or both eyes.

Further investigation of both of these features could lead to a better features set and a more accurate classification.

6.3.5 Use Color Slope To Classify Diopter Error

The two most common disorders are myopia (nearsightedness) and hyperopia (farsightedness). Myopia occurs when the eye is elongated, while hyperopia occurs when the eye is shortened. In each case, these are used as a measure of where the light focuses, either in front of or behind the retina. The medical community has established a convention for measuring the degree of focusing error using diopters. For a perfectly spherical eye, the refractive error is 0.00 diopters, while negative values indicate myopia and positive values indicate hyperopia.

The goal is to use the color slope and intensities of the reflection off the back of the pupil to accurately classify severe myopia or hyperopia. The research would look to classify the patients accurately within one diopter. One possible approach would be to classify the patients into eight groups: less than -5, between -5 and -4, between -4 and -3, between -3 and 0, between 0 and 3, between 3 and 4, between 4 and 5, and greater than 5.

Success in this research area would further advance the utility of the AVVDA system and allow the operator to refer the patient to a specialist and get fitted for corrective lenses. In addition it would act as a preventative measure for amblyopia as refractive error is considered one of the causes of amblyopia.

6.3.6 Conclusion

The research presented in this paper focused on improving the performance of an existing system of identifying patients that have a high risk of developing the amblyopic condition and should be referred to a specialist. The existing system used artificial intelligence classifiers on a set of fifty six features. These features were first defined as candidates from the research presented by Dr. Cibis in a 1994 paper that documented a manual process that classified amblyopic factors identified from patient videos [Cibis 1994]. The process was fairly meticulous and required a specialist to both collect and analyze the data extracted from the images [Cibis 1994]. Wang and Van Eenwyk expanded on the system by automating the image processing, feature extraction, and classification step through the work accomplished with the AVVDA system [Wang 2005][Van Eenwyk 2008]. The AVVDA system was able to accomplished these goals with a 77% accuracy. The work presented in this paper investigates the use of a new feature set derived from the feature extraction techniques used by Wang. The primary focus was the pixel color across several axes and the rate of change of that color across the iris. After processing and testing the data using a 10-fold stratified cross validation procedure, the best result achieved an overall accuracy of 68%. While the results did not outperform the existing system; the process allowed a thorough examination of color slope as a potentially feature for identifying amblyopia. In addition, during the investigation of the color slope feature set, there were additional research opportunities identified that hold the potential to further advance the accuracy of the AVVDA system.

References

[Anderson 1999] S. J. Anderson, I. E. Holliday, & G. F. Harding “Assessment of cortical dysfunction in human strabismic amblyopia using magnetoencephalography,” (MEG). *Vision Research*, vol. 39, pp. 1723–1738, 1999.

[Breiman 2001] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5-32, 2001.

[Cibis 1994] G. W. Cibis, "Video Vision Development Assessment (VVDA): Combining the Bruckner Test with Eccentric Photoreflection for Dynamic Identification of Amblyogenic Factors," *Transactions of the American Ophthalmological Society*, vol. 84, pp. 643-685, 1994.

[Cibis 2005] G. W. Cibis, "Video vision development assessment in diagnosis and documentation of microtropia," *Binocular Vision & Strabismus Quarterly*, vol. 20, pp. 151-158, 2005.

[Freedman 1992] H. L. Freedman and K. L. Preston, "Polaroid Photoscreening for Amblyogenic Factors. An Improved Methodology," *Ophthalmology*, vol. 99, pp. 1785-1795, 1992.

[Group 2005] The Vision in Preschoolers Study Group, "Preschool Vision Screening Tests Administered by Nurse Screeners Compared with Lay Screeners in the Vision in Preschoolers Study," *Investigative Ophthalmology & Visual Science*, vol. 46, pp. 2639-2648, 2005.

[Kam Ho 1995] T. K. Ho, “Random Decision Forest,” *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278-282, 1995.

[Kemper 2007] A. R. Kemper, P. A. Margolis, S. M. Downs, and W. C. Bordley, “A Systematic Review of Vision Screening Tests for the Detection of Amblyopia,” *Pediatrics*, vol. 104, pp. 1220-1222, 2007.

[Kohavi 1995] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *International Joint Conference and Artificial Intelligence*, pp. 1134-1143, 1995.

[Mathworks] The MathWorks Inc., "MATLAB," Available at HTTP: http://www.mathworks.com/products/matlab/

[Pediatrics 2002] American Academy of Pediatrics, "Use of photoscreening for children’s vision screening," *Pediatrics*, vol. 109, pp. 524-525, 2002.

[Pao 1989] Y. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Company, 1989.

[Quinlan 1993] J. R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1993.

[Robaei 2006] D. Robaei, K. Rose, E. Ojaimi, A. Kifley, F. Martin, and P. Mitchell, "Causes and Associations of Amblyopia in a Population-Based Sample of 6-Year-Old Australian Children," *Archives of Ophthalmology*, vol. 126, pp. 878-884, 2006.

[Russell 1995] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall Inc., 1995.

[Sheridan 1973] Sheridan MD., "The STYCAR graded-balls vision test," *Developmental Medicine and Child Neurology*, vol. 15, pp. 423-432, 1973.

[Steinman 2000] S. B. Steinman, B. A. Steinman, and R. P. Garzia, *Foundations of Binocular Vision: A Clinical Perspective*, McGraw-Hill, 2000.

[Van Eenwyk 2008] J. Van Eenwyk, A. Agah, and G. Cibis, "Automated Human Vision Assessment Using Computer Vision and Artificial Intelligence," *Proceedings of the IEEE international Conference on System of Systems Engineering (SoSE 2008)*, Monterey, California, June 2008.

[Wang 2002] T. Wang, "Eye Location and Fixation Estimation Techniques for Automated Video Vision Development Assessment," in *Computer Engineering and Computer Science*: University of Missouri - Columbia, 2002.

[Wang 2005] T. Wang, "Investigation of Image Processing and Computer-Assisted Diagnosis System for Automated Video Vision Developement Assessment," in *Computer Engineering and Computer Science Columbia*: University of Missouri - Columbia, 2005.

[Weber 2005] J. L. Weber and J. Wood, "Amblyopia: Prevalence, Natural History, Functional Effects and Treatment," *Clinical and Experimental Optometry*, vol. 88, pp. 365-375, 2005.

[Wheeler 1942] M. Wheeler, "Objective Strabismometry in Young Children," *Trans Am Opthalmol Society*, vol. 40, pp. 547-564, 1942.

[Witten 2005] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition: Morgan Kaufmann, San Francisco, 2005.